

PATENT ABSTRACTS OF JAPAN

(11) Publication number : 10-254875
 (43) Date of publication of application : 25. 09. 1998

(51) Int. Cl.

G06F 17/27

(21) Application number : 09-057283
 (22) Date of filing : 12. 03. 1997

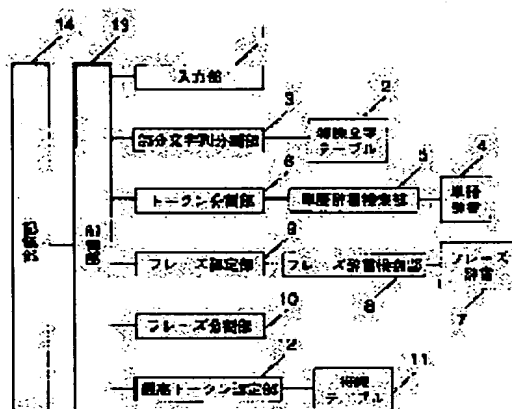
(71) Applicant : MATSUSHITA ELECTRIC IND CO LTD
 (72) Inventor : KINOSHITA HITOMI

(54) JAPANESE MORPHEME ANALYSIS DEVICE AND METHOD

(57) Abstract:

PROBLEM TO BE SOLVED: To eliminate a wasteful processing that a dictionary is retrieved by a retrieval key containing a special character and to improve processing speed by dividing an inputted character string into partial character strings through the use of a special character table and dividing the divided partial character strings into tokens by referring to a word dictionary.

SOLUTION: The Japanese character string is inputted from an input part 1. A partial character string division part 3 divides an input sentence into the partial character strings with the special character registered in a special character table 2 as a delimiter. Then, the partial character strings are divided into the tokens by using the word dictionary 4. The character registered in the special character table 2 is not contained in the index of the word dictionary 4. A word dictionary retrieval part 5 retrieves the word dictionary 4 with the Japanese character string as the key. A token division part 6 segments the token by using the result of the word dictionary retrieval part 5. Thus, the wasteful processing that the dictionary is retrieved by the retrieval key containing the special character is eliminated and the processing speed of morpheme analysis is improved.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C) ; 1998, 2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平10-254875

(43)公開日 平成10年(1998)9月25日

(51) Int.Cl.⁸
G 0 6 F 17/27

識別記号

F I
G 0 6 F 15/38

E

審査請求 未請求 請求項の数9 O.L (全 7 頁)

(21)出願番号 特願平9-57283

(22)出願日 平成9年(1997)3月12日

(71)出願人 000005821

松下電器産業株式会社

大阪府門真市大字門真1006番地

(72)発明者 木下ひとみ

大阪府門真市大字門真1006番地 松下電器
産業株式会社内

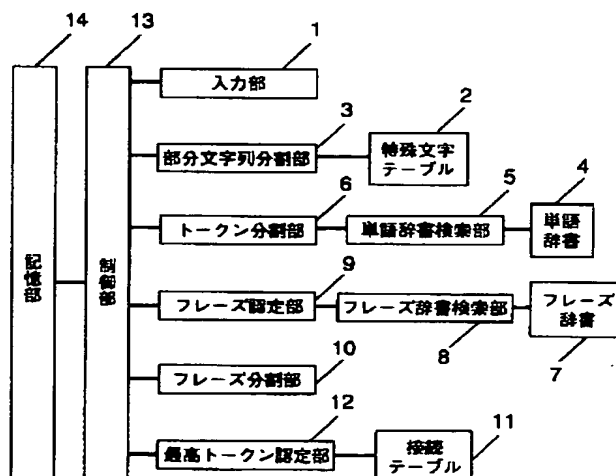
(74)代理人 弁理士 滝本 智之 (外1名)

(54) 【発明の名称】 日本語形態素解析装置及び日本語形態素解析方法

(57) 【要約】

【課題】 特殊文字を含む検索キーで辞書検索するという無駄な処理を省き、処理速度を向上させることのできる日本語形態素解析装置及び日本語形態素解析方法を提供すること。

【解決手段】 日本語文を文字列として入力する入力部 1 と、日本語単語の読み、漢字表記、品詞など、形態素解析に必要な語彙情報を記憶した単語辞書 4 と、日本語文字列をキーとして単語辞書 4 を検索する単語辞書検索部 5 と、句読点や括弧など前記辞書の登録見出しに含まれない文字（特殊文字）を定義した特殊文字テーブル 2 と、入力部 1 より入力された文字列を特殊文字で部分文字列に分割する部分文字列分割部 3 と、部分文字列分割部 3 で分割した部分文字を単語辞書検索部 5 を用いて単語辞書 4 を参照し、トークンに分割するトークン分割部 6 とを有する。



【特許請求の範囲】

【請求項1】日本語文字列を入力する入力手段と、句読点や括弧などの前記単語辞書の登録見出しに含まれない特殊文字を記憶した特殊文字テーブルと、前記入力手段より入力された文字列を、前記特殊文字テーブルを用いて部分文字列に分割する部分文字列分割手段と、

日本語単語の読み、漢字表記、品詞などの形態素解析に必要な語彙情報を記憶した単語辞書と、

日本語文字列をキーとして前記単語辞書を検索する辞書検索手段と、

前記部分文字列分割手段で分割した部分文字列を、前記辞書検索手段を用いて前記単語辞書を参照し、トークンに分割するトークン分割手段と、を有することを特徴とする日本語形態素解析装置。

【請求項2】登録見出しに特殊文字を含むフレーズ辞書と、

前記入力手段より入力される入力文字列を前記フレーズ辞書を用いて辞書検索を行うフレーズ辞書検索手段と、前記フレーズ辞書検索手段を用いて前記フレーズ辞書を参照し、フレーズを認定するフレーズ認定手段と、を有し、

前記フレーズ認定手段によりフレーズ認定し、認定できなかった部分に対して、前記部分文字列分割手段は分割を行うことを特徴とする請求項1記載の日本語形態素解析装置。

【請求項3】前記フレーズ認定手段により認定されたフレーズについて、前記辞書検索手段を用いて前記単語辞書を参照し、トークンに分割するフレーズ分割手段と、隣接するトークンの接続可否を定義した接続テーブルと、

前記フレーズ認定手段の認定結果と前記フレーズ分割手段の分割結果とを照合し、前記接続テーブルに基づいていずれか一方を決定するトークン認定手段と、を有することを特徴とする請求項2記載の日本語形態素解析装置。

【請求項4】データを入力する入力装置と、データを記憶する記憶装置と、を有する装置を制御する日本語形態素解析方法であって、

日本語文を文字列として入力するステップと、入力された文字列を、特殊文字テーブルを用いて部分文字列に分割するステップと、分割した部分文字列を、単語辞書を参照して辞書検索し、トークンに分割するステップと、を有することを特徴とする日本語形態素解析方法。

【請求項5】文字列が入力された後、フレーズ辞書を参照して辞書検索し、フレーズの認定を行うステップを有し、認定できなかった部分に対して部分文字列に分割することを特徴とする請求項4記載の日本語形態素解析方法。

【請求項6】フレーズを認定した後、認定されたフレーズについて、単語辞書を参照して辞書検索し、トークンに分割するステップと、

フレーズの認定結果と認定したフレーズの分割結果とを照合し、接続テーブルに基づいていずれか一方を決定するステップと、を有することを特徴とする請求項5記載の日本語形態素解析方法。

【請求項7】データを入力する入力装置と、データを記憶する記憶装置と、を有する装置を制御するためのプログラムを記憶した記憶媒体であって、

日本語文を文字列として入力するステップと、入力された文字列を、特殊文字テーブルを用いて部分文字列に分割するステップと、

分割した部分文字列を、単語辞書を参照して辞書検索し、トークンに分割するステップと、を有することを特徴とするプログラムを記憶した記憶媒体。

【請求項8】文字列が入力された後、フレーズ辞書を参照して辞書検索し、フレーズの認定を行うステップを有し、認定できなかった部分に対して部分文字列に分割することを特徴とする請求項7記載のプログラムを記憶した記憶媒体。

【請求項9】フレーズを認定した後、認定されたフレーズについて、単語辞書を参照して辞書検索し、トークンに分割するステップと、

フレーズの認定結果と認定したフレーズの分割結果とを照合し、接続テーブルに基づいていずれか一方を決定するステップと、を有することを特徴とする請求項7記載のプログラムを記憶した記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文字列として入力した日本語文の形態情報を出力する日本語形態素解析装置及び日本語形態素解析方法に関するものである。

【0002】

【従来の技術】例えば仮名漢字変換や機械翻訳など、日本語の処理を実行するにあたっては、形態素解析と呼ばれる処理を行う必要がある。

【0003】従来、形態素解析は、単語をキーとしてその語彙情報を記憶した辞書を検索しながら、文字列を形態素（意味を持つ最小の単位であり、以下「トークン」と称する）に分割し（以下「トークン分割」と称する）、個々のトークンに対して形態情報（品詞、活用など）を付加していた。

【0004】日本語の形態素解析における辞書検索では、一般に次の2通りの手法が考えられている。まず1つは、入力文の任意の位置から始まるもっとも長い文字列（トークン）を獲得する方法で、最長一致の考え方である。しかし、最長一致法だけでは誤解釈を招くため、もう1つの手法として、入力文の任意の位置から始まる全ての可能なトークンを獲得し、接続テーブルやトーク

10

20

30

40

50

ンの優先度などその他様々な情報を用いて、最適なトークン認定を行うものがより一般化している。

【0005】

【発明が解決しようとする課題】しかしながら、入力文は句読点や括弧などの特殊文字を含んでおり、前述の辞書検索手法のいずれも、この特殊文字を含む文字列を検索キーとしてしまうという問題点を有している。

【0006】例えば、入力文が『昨日、公園で彼女に会った。』であった場合、『昨日、』のように読点を含む文字列で辞書を検索してしまう。日本語の形態素解析処理では、辞書検索の占める割合が高く、辞書検索の処理速度は日本語の形態素解析処理の速度に多大な影響を及ぼすにも関わらず、従来手法では、特殊文字を含む文字列で辞書を検索してしまうという無駄な処理を行っている。

【0007】本発明は、特殊文字を含む検索キーで辞書検索するという無駄な処理を省き、処理速度を向上させることのできる日本語形態素解析装置及び日本語形態素解析方法を提供することを目的とする。

【0008】

【課題を解決するための手段】この課題を解決するために本発明は、日本語文字列を入力する入力手段と、句読点や括弧などの前記単語辞書の登録見出しに含まれない特殊文字を記憶した特殊文字テーブルと、入力手段より入力された文字列を、特殊文字テーブルを用いて部分文字列に分割する部分文字列分割手段と、日本語単語の読み、漢字表記、品詞などの形態素解析に必要な語彙情報を記憶した単語辞書と、日本語文字列をキーとして前記単語辞書を検索する辞書検索手段と、部分文字列分割手段で分割した部分文字列を、辞書検索手段を用いて単語辞書を参照し、トークンに分割するトークン分割手段と、を有する構成としたものである。

【0009】これにより、特殊文字を含む検索キーで辞書検索するという無駄な処理を省き、形態素解析の処理速度を向上させることのできる日本語形態素解析装置が得られる。

【0010】

【発明の実施の形態】本発明の請求項1に記載の発明は、日本語文字列を入力する入力手段と、句読点や括弧などの前記単語辞書の登録見出しに含まれない特殊文字を記憶した特殊文字テーブルと、入力手段より入力された文字列を、特殊文字テーブルを用いて部分文字列に分割する部分文字列分割手段と、日本語単語の読み、漢字表記、品詞などの形態素解析に必要な語彙情報を記憶した単語辞書と、日本語文字列をキーとして前記単語辞書を検索する辞書検索手段と、部分文字列分割手段で分割した部分文字列を、辞書検索手段を用いて単語辞書を参照し、トークンに分割するトークン分割手段と、を有する構成としたものであり、特殊文字を含む検索キーで辞書検索するという無駄な処理を省くことができる。

10

20

30

40

50

【0011】本発明の請求項2に記載の発明は、請求項1に記載の発明において、登録見出しに特殊文字を含むフレーズ辞書と、入力手段より入力される入力文字列をフレーズ辞書を用いて辞書検索を行うフレーズ辞書検索手段と、フレーズ辞書検索手段を用いてフレーズ辞書を参照し、フレーズを認定するフレーズ認定手段と、を有し、フレーズ認定手段によりフレーズ認定し、認定できなかった部分に対して、部分文字列分割手段は分割を行う構成としたことにより、特殊文字を含むトークンの定義も可能にし、さらに、辞書に存在しないことが明らかな特殊文字を含む検索キーで辞書検索するという無駄な処理を省くことができる。

【0012】本発明の請求項3に記載の発明は、請求項2に記載の発明において、フレーズ認定手段により認定されたフレーズについて、辞書検索手段を用いて単語辞書を参照し、トークンに分割するフレーズ分割手段と、隣接するトークンの接続可否を定義した接続テーブルと、フレーズ認定手段の認定結果とフレーズ分割手段の分割結果とを照合し、接続テーブルに基づいていずれか一方を決定するトークン認定手段と、を有する構成としたものであり、単語及びフレーズのうち、最適な解釈を獲得できる。

【0013】

【実施例】以下、本発明の実施例について、図面を参照しながら説明する。

【0014】図1は、本発明の一実施例における日本語形態素解析装置の機能ブロック図である。

【0015】図1において、1は操作者から日本語文字列のデータが入力される入力部である。2は句読点や括弧など単語辞書に登録されない特殊文字を登録した特殊文字テーブルである。本実施例では図3に示すように、特殊文字テーブル2には、26種類の文字を特殊文字として定義されている。3は特殊文字テーブル2に登録された文字を区切り文字として、入力文字列を部分文字列に分割する部分文字列分割部である。

【0016】4は文字列をキーとして、その語彙情報が登録された単語辞書である。単語辞書4の見出しには、特殊文字テーブル2に登録された文字は含まない。5は日本語文字列をキーとして、単語辞書4を検索する単語辞書検索部である。6は単語辞書検索部5の結果を用いてトークンを切り出すトークン分割部である。

【0017】7は特殊文字テーブル2に登録された特殊文字を含む文字列をキーとして、その語彙情報が登録されたフレーズ辞書である。フレーズ辞書7及び単語辞書4に登録された見出しには、それぞれ優先度が付加されている。8は日本語文字列をキーとして、フレーズ辞書7を検索するフレーズ辞書検索部である。9はフレーズ辞書検索部8の結果を用いてフレーズを認定するフレーズ認定部である。10はフレーズ認定部9で認定されたフレーズを単語辞書4を用いて再分割するフレーズ分割

部である。

【0018】11は隣接するトークンの接続可否を定義した接続テーブルである。ここで用いる接続テーブルは、行を後接情報、列を前接情報とした2次元テーブルであり、接続不可の場合は0、接続可の場合は接続関係の強さによって1から5の数値が予め設定されている（接続関係がもっとも強いものは1とする）。12は接続テーブル11、単語辞書4及びフレーズ辞書7より取得した見出しの優先度を基に最適なトークンを決定する最適トークン認定部である。

【0019】13は入力部1、部分文字列分割部3、トークン分割部6、フレーズ認定部9、フレーズ分割部10、最適トークン認定部12を制御し、入力部1で入力された文の形態情報を出力する制御部である。

【0020】14は入力部1で入力された文、部分文字列分割部3の分割結果、単語辞書検索部5の検索結果、トークン分割部6の分割結果、フレーズ辞書検索部8の検索結果、フレーズ認定部9の認定結果、フレーズ分割部10の分割結果、最適トークン認定部12の認定結果を記憶する記憶部である。

【0021】図2は本発明の一実施例における日本語形態素解析装置の回路ブロック図である。

【0022】図2において、21はキーボードやマウスなどの入力装置である。22は陰極線管ディスプレイ（CRT）などの表示装置である。23は中央処理装置（CPU）である。24はランダム・アクセス・メモリ（RAM）である。25はリード・オンリー・メモリ（ROM）である。

【0023】図1及び図2に示すように、入力部1は入力装置21により実現される。記憶部14はRAM24により実現される。特殊文字テーブル2及び接続テーブル11はROM25により実現される。単語辞書4及びフレーズ辞書7はRAM24、ROM25、または2次記憶装置（図示せず）のいずれかに記憶されている。

【0024】また、部分文字列分割部3、単語辞書検索部5、トークン分割部6、フレーズ辞書検索部8、フレーズ認定部9、フレーズ分割部10、最適トークン認定部12及び制御部13は、CPU23がRAM24及びROM25とデータのやりとりを行いながら、ROM25に記憶されたプログラムを実行することにより実現される。

【0025】以上のように構成された日本語形態素解析装置について、以下にその動作を説明する。なお、図4及び図6に示すフローチャートは、CPU23がROM25に記憶されているプログラムを実行する様子を示したものである。

【0026】図4は本発明の一実施例における日本語形態素解析処理のフローチャートであり、入力文を特殊文字で部分文字列に分割し、個々の部分文字列に対して、単語辞書4を用いて形態素解析する処理の流れを示した

ものである。

【0027】図4に示すように、ステップ11では、入力部1より日本語文字列が入力される。なお、以降で行う文字列操作は、全て全角文字を対象とする。従って、ここで半角文字が入力された場合は全角に変換しておく。また、ここでは次の入力文aが与えられたものとする。

【0028】（a）昨日、私は「教授」と呼ばれる人に出会った。が与えられたものとする。

10 【0029】次に、ステップ12では、部分文字列分割部3によって、特殊文字テーブル2（図3参照）に登録された特殊文字を区切り文字として、入力文aを部分文字列に分割する。分割した部分文字列の数は、ステップ13において、変数Nとして格納される。

【0030】ステップ13における分割後の入力文aの構造を図5に示す。図5に示すように、入力文aは、部分文字列と区切り文字となった特殊文字のリスト構造となり、特殊文字に該当する部分には、その旨を示すフラグを立てる。図5の記号Sはこのリストの先頭ポインタを示している。また、変数Nの値は部分文字列と区切り文字の総数で、入力文aでは8となる。

20 【0031】ステップ14では、カウンタ変数iに0を、ポインタ変数pにSをそれぞれセットする。ステップ15では、カウンタ変数iと変数Nとの値を比較し、変数Nの方が大きければステップ16に進み、そうでなければステップ19に進む。

【0032】ステップ16では、ポインタ変数pが指す文字列データが、特殊文字であるのか否かチェックし、特殊文字であればステップ18に進み、特殊文字でなければステップ17に進む。

30 【0033】ステップ17では、ポインタ変数pが指す文字列データを、単語辞書4を用いてトークンに分割する。ここで、トークン分割の処理の流れを図6のフローチャートに基づいて説明する。

【0034】図6は本発明の一実施例における日本語形態素解析処理のフローチャートであり、辞書見出しと一致する最長のトークンを切り出す（最長一致）方法でトークン分割を行うものとする。なお、ポインタ変数pは、文字列データ『と呼ばれる人に出会った』を指しているものとする。

40 【0035】まず、ステップ21では、変数Lenに入力文字列の長さ（バイト数、この例では22となる）を、ステップ22で変数posに0をセットする。

【0036】ステップ23では、変数posと変数Lenとを比較し、変数posの方が小さければステップ24に進み、そうでなければトークン分割の処理を終えて図4に戻る。

50 【0037】ステップ24では、変数keyLenに“Len-pos”（現時点で最長の文字数）の値をセットする。ステップ25では、変数keyLenが0よ

り大きいかな否かチェックし、大きければステップ26に進み、そうでなければステップ2Aに進む。

【0038】ステップ26では、変数keyに入力文字列の変数pos番目から変数Lenバイトの文字を格納し、変数keyの文字列を検索キーとして単語辞書検索部5へ渡す。

【0039】ステップ28では、単語辞書検索部5で単語辞書4中に変数keyと一致する見出しが検索できたかな否かチェックし、検索できていればステップ2Bに進み、そうでなければステップ29に進む。

【0040】ステップ29では、変数keyは辞書に登録されていないことが判明したので、検索キーの長さを1文字分減らす。keyLenが0になるか、辞書中にマッチする語が見付かるまで、ステップ25からステップ29までの処理を繰り返す。ステップ2Aでは、変数posから始まる文字列が辞書中に存在しなかったもので、posから始まる1文字を未知語として処理する。

【0041】ステップ2Bでは、変数posから始まる文字列が辞書中に存在したので、その語をトークンとして切り出す。

【0042】ステップ2Cでは、変数posを切り出したトークンの文字数分増加させる。変数posが変数Lenより大きくなるまでステップ24からステップ2Cまでの処理を繰り返す。部分文字列単位のトークン分割処理が終了するとステップ18(図4)に戻る。

【0043】図4のステップ18では、変数iを1増やし、変数pに次の文字列データのポインタをセットする。iがNと等しくなるまでステップ16からステップ18までの処理を繰り返す。ステップ19では、形態素解析の結果を出力する。解析結果を図7に示す。

【0044】以上のように、本実施例の日本語の形態素解析が行われるのであるが、上記説明では単語辞書4のみを用いた場合について説明したものである。しかしながら、『X'mas』のように、特殊文字を含む単語を辞書に登録したい場合も考えられる。

【0045】ところが、このような場合に、『X'mas』を単語辞書に登録し、図4のフローチャートに従って処理を行うと、部分文字列分割部3は、『X'mas』は「X」、「'」、「mas」の3つの部分文字列に分割されてしまい、うまく処理できないことがある。

【0046】そこで、特殊文字を含む文字列を登録するフレーズ辞書7を別途設け、入力部1から文字列が入力されると、まず、フレーズ辞書7を用いてフレーズ辞書検索部8により検索し、フレーズ認定部9により入力文に含まれるフレーズを検出し、この部分に関しては図4で示した処理を行わないようにする。これにより、

『X'mas』をうまく処理できるようになる。

【0047】ところが、フレーズ辞書7を用いて『X'mas』の問題を解決できるのであるが、実は、前後の文字によっては、「X」、「'」、「mas」と分割し

た方が良い場合も考えられる。そのような場合のため、フレーズ認定部9によってフレーズ認定された文字列を今度は単語辞書4を用いて、再分割を行うフレーズ分割部10を設け、フレーズ認定部9の認定結果とフレーズ分割部10の分割結果とのうち、接続テーブル11に基づいて、前後のトークンとの接続の強度やトークンの優先度から考慮して最適な一方を決定する最適トークン認定部12を設けると良い。

【0048】以上のように本実施例では、特殊文字を含む文字列のように、辞書に登録されていないことが明らかな検索キーで辞書検索するという無駄な処理を省き、形態素解析の処理速度を向上させることができる。

【0049】また、特殊文字を含む文字列を見出しとするフレーズ辞書を別途用意することで、本発明の利点を残したまま、フレーズに対応することが可能となる。

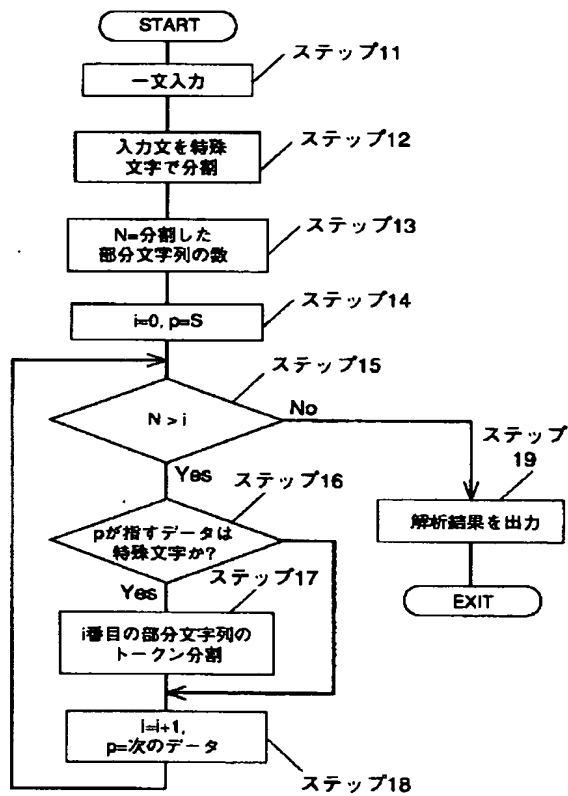
【0050】

【発明の効果】以上のように本発明によれば、日本語文字列を入力する入力手段と、句読点や括弧などの前記単語辞書の登録見出しに含まれない特殊文字を記憶した特殊文字テーブルと、入力手段より入力された文字列を、特殊文字テーブルを用いて部分文字列に分割する部分文字列分割手段と、日本語単語の読み、漢字表記、品詞などの形態素解析に必要な語彙情報を記憶した単語辞書と、日本語文字列をキーとして前記単語辞書を検索する辞書検索手段と、部分文字列分割手段で分割した部分文字列を、辞書検索手段を用いて単語辞書を参照し、トークンに分割するトークン分割手段と、を有する構成としたものであり、特殊文字を含む検索キーで辞書検索するという無駄な処理を省くことができるので、日本語の形態素解析の処理速度を大幅に向上することができる。

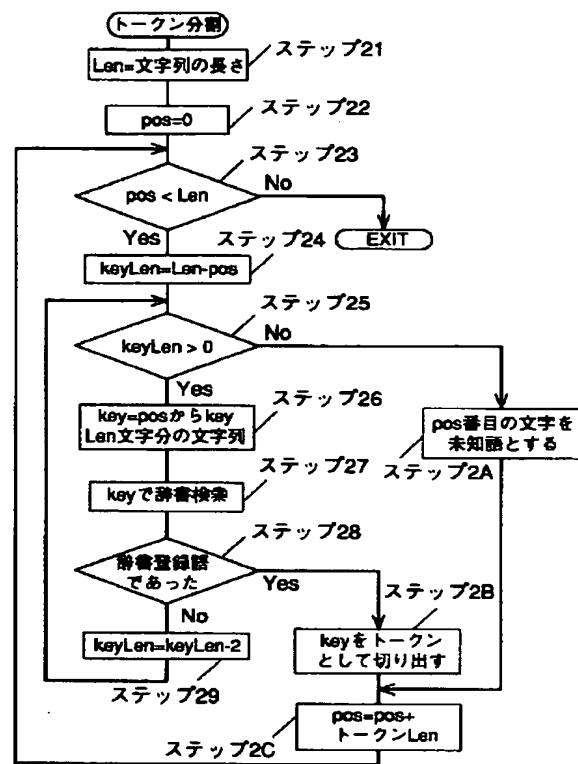
【0051】また、登録見出しに特殊文字を含むフレーズ辞書と、入力手段より入力される入力文字列をフレーズ辞書を用いて辞書検索を行うフレーズ辞書検索手段と、フレーズ辞書検索手段を用いてフレーズ辞書を参照し、フレーズを認定するフレーズ認定手段と、を有し、フレーズ認定手段によりフレーズ認定し、認定できなかった部分に対して、部分文字列分割手段は分割を行う構成としたことにより、特殊文字を含むトークンの定義も可能にし、さらに、辞書に存在しないことが明らかな特殊文字を含む検索キーで辞書検索するという無駄な処理を省くことができるので、特殊文字を含むトークンの分割の精度がより向上する。

【0052】さらに、フレーズ認定手段により認定されたフレーズについて、辞書検索手段を用いて単語辞書を参照し、トークンに分割するフレーズ分割手段と、隣接するトークンの接続可否を定義した接続テーブルと、フレーズ認定手段の認定結果とフレーズ分割手段の分割結果とを照合し、接続テーブルに基づいていずれか一方を決定するトークン認定手段と、を有する構成としたものであり、単語及びフレーズのうち、最適な解釈を獲得で

【図4】



【図6】



【図7】

